
Research article

A comparison of predictions for cardiovascular disease using a variety of machine learning techniques

Khamis K. Al-karawi^{1, 2*}, Abdullah K. Al-karawi³

¹ School of Science, Engineering, and Environment, Salford University, Greater Manchester, UK, k.a.yousif@edu.salford.ac.uk

² University of Diyala, Baqubah, Diyala, Iraq

³ Arden University, Greater Manchester, UK

* Correspondence: alkasi_68@yahoo.com

Keywords:

SVM, Random Forest, Decision tree, logistic regression, Machine learning, Cardiovascular disease prediction

Important Dates:

Received: 8 June 2026

Revised: 30 June 2026

Accepted: 6 July 2026

Online: 15 July 2026

Copyright © 2026 by the authors. Published under Creative Commons Attribution (CC BY) license.

ABSTRACT

One of the leading causes of global death is cardiovascular disease. They also pose a significant threat to people in general. Early access to heart disease detection is critical for increasing survival rates. There are multiple factors, such as age, gender, cholesterol, blood sugar, and heart rate, that can influence the risk of suffering from a life-threatening problem. However, it is difficult for experts to assess these variables when predicting a patient's risk profile. This study employs machine learning to estimate the likelihood of disease occurrence in patients. In this work, the authors propose applying those learning's using the UCI Machine Learning Heart Diagnosis dataset and various machine learning algorithms. This study aims to compare, contrast, and evaluate the results of machine learning algorithms on the UCI Machine Learning Heart Disease dataset. The study also aims to determine which machine learning algorithms are well-suited for detecting heart disease. It is well known that computational methods can be a promising platform for disease prediction from the UCI Machine Learning Heart Diagnosis dataset. The commonly used machine learning algorithms include kernel configuration only (SVM), random forest, and decision tree, and additional machine learning algorithms. Each method offers substantial potential for disease prediction, making them helpful in treating cardiovascular diseases in general. We found that random forests are simple, and you can get good results with around 87.76% accuracy.

1. Introduction

Cardiovascular disease (CVD) continues to be responsible for more deaths globally, with death rates reaching significantly higher levels every year. Early prediction of cardiovascular disease represents an opportunity to prevent a large number of deaths. The integration of artificial intelligence is a new approach to predicting cardiovascular diseases. Different machine learning methods have been employed to capture hospital patient details to perform better and quicker predictions. Many are also identified as heart disease risk factors. Such risk factors may be based on age, sex, genetic components, and additional health elements such as high blood pressure and diabetes [1, 2]. Cardiovascular disease (CVD) is the leading cause of death worldwide, and the death rate is increasing significantly every year. An early prediction system is an excellent basis for preventing most deaths. Artificial intelligence has recently been used in this field since this approach has already increased the predictability levels. Many machine learning algorithms have already been implemented in data prediction. Artificial intelligence systems can readily acquire hospital patient data for more accurate and faster forecasts. Many additional

variables have been discovered as risk factors for heart disease [2, 3]. This might include the patient's age, gender, genetics, and other health conditions, including high blood pressure and diabetes. According to the World Health Organization (WHO), many people succumb to various types of cardiovascular events, including heart attacks and chest discomfort, each year [4, 5]. Heart disease is a broad phrase that encompasses several heart-related disorders. According to World Health Organization figures, cardiovascular disorders are now the top cause of mortality globally, killing 17.9 million people each year [6]. Engaging in harmful behaviors, such as high cholesterol, obesity, raised triglyceride levels, hypertension, and others, considerably increases the chance of developing heart disease. The American Heart Association emphasizes specific indications[7], including sleep difficulties, abnormal heart rate (arrhythmia), swollen legs, and fast weight gain, which can exceed 1-2 kg daily [8, 9]. These symptoms may match those of other disorders, such as those observed in older people, making proper diagnosis difficult and potentially deadly.

Over time, research data and hospital patient information have become available. Various open sources provide access to patient information, allowing researchers to undertake studies that use computer technology to diagnose and identify illnesses accurately before they become deadly. Machine learning and artificial intelligence play a significant role in the medical industry. We can accurately categorize or forecast outcomes and diagnose illnesses with machine learning and deep learning models. Machine learning algorithms make accurate predictions possible, making it easier to analyze genetic data in-depth. Medical records may be changed and thoroughly examined to improve predictive capacities, and these models can be taught to make pandemic predictions based on information [8, 10]. Numerous studies have explored applying machine learning models to classify and predict heart disease diagnosis.

2. Related work

The human heart's electrical activity is captured through electrocardiography (ECG), employing skin electrodes to record various waveforms. This non-invasive technique provides insights into heartbeat, heart rate, and overall cardiac health, making it instrumental in identifying heart diseases. Notably, the numerous cells in the human body do not have direct external contact; instead, they rely on the cardiovascular system for transportation. The cardiovascular system comprises two main fluid types: blood and lymph. Blood, coursing through blood vessels and the heart, constitutes the first type. Lymph is the second kind; lymphatic arteries and lymph nodes are parts of its structure. The cardiovascular system comprises the lymphatic and vascular systems working together [11, 12]. A cardiac cycle is a series of events that make up a heartbeat. A cardiac cycle usually consists of the contraction of both ventricles in unison with the contraction of both atria a short while later. Heart muscle cells are linked, so when one contracts, it stimulates neighboring cells. Intervals in the cardiac cycle cause the muscles to relax, aiding aerobic breathing. Even though earlier research has shown encouraging outcomes in automated cardiovascular disease (CVD) detection, several obstacles still need to be addressed. A significant issue is the unpredictability and small size of private datasets, especially in magnetocardiogram (MCG) investigations with fewer larger public datasets.

On the other hand, public datasets from research, including electrocardiograms (ECGs), could not perform as well when moving from benchmark to clinical domains. Public ECG datasets may show biases towards cases of abnormality that are easily recognized, which might affect early diagnosis [13]. The selection of features has a significant impact on the performance of models. Previous research has included features from three primary categories: (a) information theory features, (b) time domain features, and (c) frequency domain features. Heart disease prediction models (HDPM) perform better because of the potential for machine learning (ML) models to improve heart disease diagnosis. Several research studies demonstrate this. Notably, researchers frequently use publicly accessible datasets like Cleveland and Statlog to compare the accuracy of the predictive algorithms. The

healthcare industry recently began using machine learning (ML) for clinical decision-making. Machine learning advances highlight the benefits of discriminative classifiers for the automated identification of heart disease. Prior research demonstrates the successful utilization of ML algorithms, including SVM, RF (random forest), LR (logistic regression), BPNN (back propagation neural network), and MLP (multilayer perceptron), as practical decision-making tools for predicting heart disease based on individual information. Several studies have highlighted the superior performance of hybrid models, such as RF with a linear model, MLP, Bayes Net (BN), majority voting of NB, and RF, along with two stacked SVMs [14]. In their work, Kalia Orphanou and colleagues employed a Naïve Bayes (NB) classification model, integrating Temporal Association Rules (TARs) as a feature for heart disease diagnosis. The Naïve Bayes classifier considered periodic TARs as final features. Through the possible recurrence of each TAR pattern with relevant patient history, the model achieved an accuracy of 82%.

Additionally, Teresa Princy et al. conducted a survey on various machine learning techniques for predicting the risk of heart diseases based on attributes such as gender, age, cholesterol, and pulse rate. They enhanced accuracy and attribute analysis using the K-nearest neighbor algorithm, Naïve Bayes, and neural networks. The accuracy was improved with fewer attributes, employing diverse methods [15]. Prerana and colleagues used machine learning techniques to predict the risk level using probabilistic analysis and classification (PAC) along with heart disease. They utilized map-reduce programming to handle patient records and analyze various machine learning algorithms, presenting the results graphically. This approach, available in the cloud and globally accessible, holds the potential for extending its application to diagnose different diseases such as cancer, diabetes, and brain tumors. In a separate study, Shadab et al. employed NB data mining to help users obtain answers to predefined questions in a web-based application. Doctors use these intelligent decisions to diagnose heart disease, and the accuracy of the NB algorithm can be enhanced through various techniques [16].

3. Machine Learning Algorithms

This research employed logistic regression, K-NN, SVM, decision trees, and random forest algorithms. A concise overview of each algorithm is provided below:

3.1 Logistic Regression (LR)

Logistic regression (LR) is a linear model for predicting binary variables. This technique is beneficial for classifying a new observation into an unknown group. LR serves as a distinctive method for organizing data into two mutually exclusive and exhaustive data categories [17, 18]

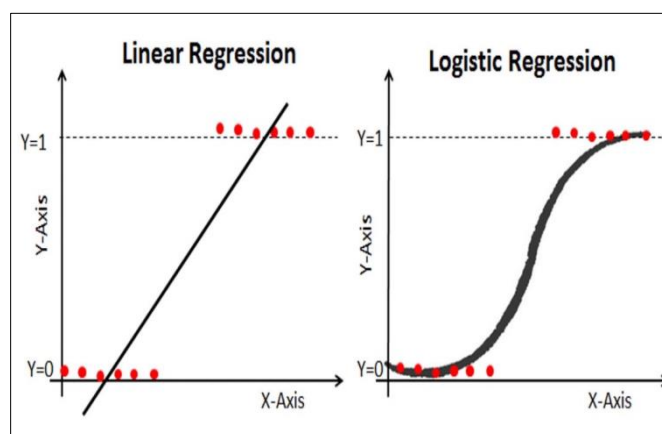


Figure.3. shows the Logistic Regression (LR)

3.2 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) represents a fundamental classification algorithm in machine learning, determining the class of an example based on the most prevalent class among its nearest neighbors within a specified parameter K [19]. In this study, we employed 47 K neighbors with uniform weights. The constant weight function ensures that all neighboring points are equally weighted. A leaf size of 30 was utilized, influencing the

implementation speed and memory requirements. The parameters " $p = 2$ " and "metric = 'Minkowski'" were set, where " p " signifies the power parameter, and 'Minkowski' denotes the default matrix.

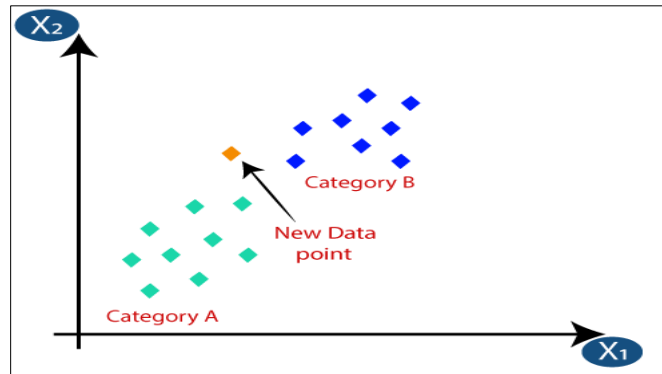


Figure.4. shows the K-Nearest Neighbors

3.3 Support Vector Machine (SVM)

The support vector machine (SVM) is an example of supervised learning. It is an effective instrument for sorting and analyzing large amounts of data. Even with vast data, the SVM classifier has a quick learning speed. SVM is used for classification problems involving two or more classes. The idea of decision planes is the conceptual foundation of the Support Vector Machine. One definition of a decision plane is dividing between collections of things belonging to different classes. Utilizing the Support Vector Machine approach allowed for the accurate classification and detection of brain tumors. The image is then classified to determine the kind of tumor in it.

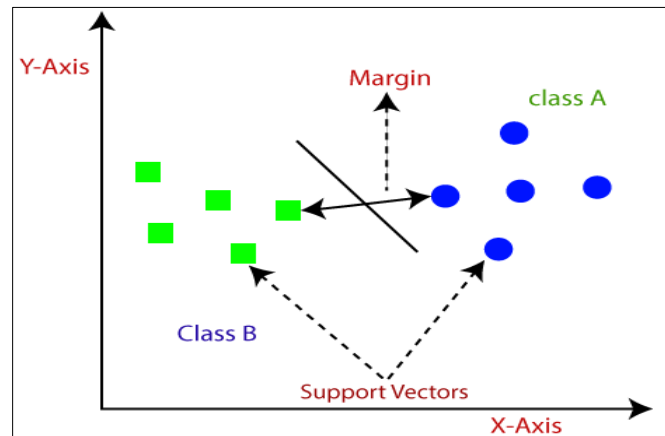


Figure 5 shows the Support Vector Machine (SVM)

Training and testing are the two fundamental processes in a support vector machine. The SVM classifier is based on the hyperplane that maximizes the separation margin between the two classes. This is true even if several hyperplanes exist for a training set that maximizes the separating margin between the two classes. This is because there may be many hyperplanes, each of which would maximize the dividing gap between the two classes. SVM classification uses a hyperplane that minimizes the number of spats between the two classes. The components of the training set that are situated on the boundary hyperplanes between the two classes are the ones that are referred to as support vectors. [16]. In this research, we utilized an "RBF" (Radial Basis Function) kernel, specifying the arrays or shapes essential for the algorithm.

3.4 Decision Tree (DT)

Decision Tree (DT) is a data exploration model that addresses classification and categorization challenges. The DT constitutes a primary recursive block, providing a cohesive framework for continuous computational implementations. It encompasses nodes (parent) and leaves (child)[20]. In this investigation, we employed three

max-leaf nodes and three criteria for Decision Trees, setting the random state to 0. The random state parameter governs the estimator's randomness, and a value of 0 indicates the absence of randomness in this context.

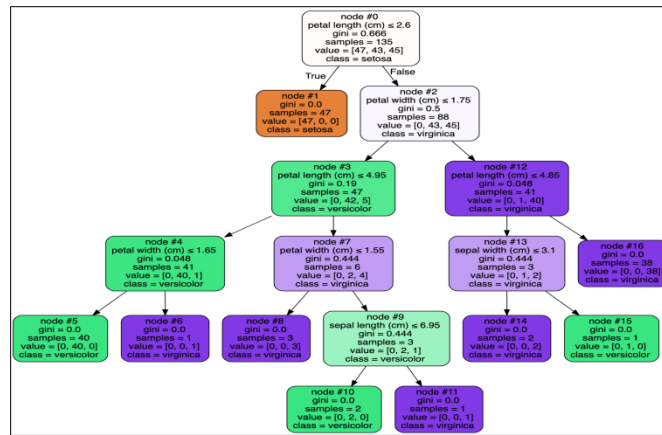


Figure.6. shows the Decision Tree

3.5 Random Forest (RF)

Random Forest (RF) stands out as a versatile and practical model with diverse outputs, effectively addressing the challenge of overfitting. It represents a prominent instance of ensemble learning, a concept involving the utilization of the same algorithm multiple times or employing various algorithms. In this research, we used 166 estimators, where estimators denote the number of trees within a forest. The "entropy" criterion, serving as a measure of quality, was utilized. The random state was set to zero and employed to control the randomness of samples during the tree-building process [21].

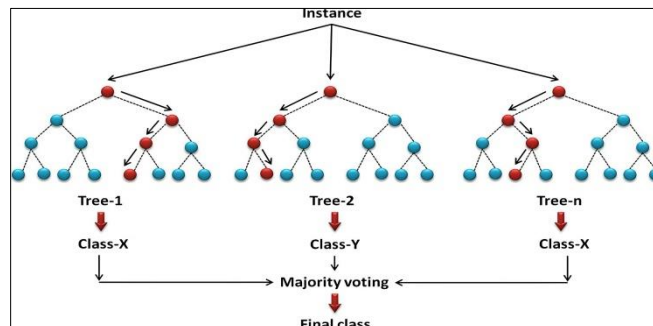


Figure.6. shows the Random Forest (RF)

4. Methodology

Early cardiac disease detection and diagnosis can significantly impact patient care and treatment. Figure 7 shows the suggested technique for gathering and processing heart disease data, including feature selection and machine learning classification. Data acquisition begins with gathering relevant data from Table 1. Pre-processing removes mistakes, missing numbers, and outliers from the gathered data and prepares it for analysis. Feature selection identifies the most relevant variables for categorization. Data normalization ensures equitable representation by aligning all values within a comparable range. The pre-processed data is separated into training and testing datasets for model training and assessment. Heart disease classification uses a variety of machine learning models, including Logistic Regression, KNN, Random Forest, Decision Tree, and Support Vector Machine (SVM)

4.1 Performance Evaluation

Following the feature selection process, the condensed dataset was divided into two segments, with 564 individuals allocated to the training data and 51 individuals forming the test dataset. The supervised learning task was executed using RF, LR, DT, K-NN, and SVM. Precision, sensitivity, specificity, and accuracy were employed

as metrics for evaluating the binary classification model. Accuracy gauges the model's proficiency in correctly identifying observations; precision evaluates its capability to discriminate between positive and negative observations; sensitivity quantifies the identification of positive classifications among all available positive observations, and specificity mirrors this interpretation for negative observations. One parameter for assessing classification models is accuracy. The percentage of detections that our model correctly detected is known as accuracy. When discussing classification based on binary data, the negative and positive calculations are used to calculate the accuracy score. The definition of accuracy is,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (1)$$

The main limitation of this evaluative measure is that it becomes imperfect when it comes in contact with a class dataset that is not balanced. Therefore, different evaluative measures are adopted to properly evaluate all the techniques used. Precision is the proportion of correctly predicted positive observations to all expected positive observations, on the other hand. A low False Positive Rate (FPR) is correlated with high precision

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

The gathered information aids in determining whether the information is connected to normal or pathological traits.

$$Sensitivity = \frac{TP}{TP + FN} * 100 \quad (3)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

In the mentioned formulas (i.e., accuracy, precision, sensitivity, specificity, and F1) and the confusion matrix, TPs denote true positives, TNs refer to true negatives, FPs signify false positives, and FNs represent false negatives. The confusion matrix provides insights into the percentages of accurate and inaccurate classifications for each class, pinpointing precisely which classes the algorithms encounter the most challenges in classification for the trained models. TP and TN signify the number of data points belonging to the positive and negative classes, respectively, where the model correctly identifies them. Conversely, FP indicates the number of negatives erroneously classified as positives, and FN represents the count of positives mistakenly classified as negatives by the machine. This matrix type evaluates different models using different machine-learning techniques or algorithms. The matrix can be represented as the representation of actual values against predicted values. The structure of the matrix can be referred to as a table. There are 4 elements when considering the confusion matrix. The first element refers to true positives, where specific values are detected as positive and are positive in actuality. Similarly, true negative, false negative, and false positive are referred to as specific values where the detected and actual values are both negative and particular, detected as positive values but are negative. There exist specific values that are negative but predicted as positive values. There are certain terms related to the matrix mentioned above:

- **Recall:** It refers to the terminology where the ratio of the values detected as positive correctly to the overall number of values present.
- **Precision** refers to the terminology that refers to the ratio of the values detected as positive and positive, to the overall positive predicted values.

5. Experiment setup

5.1 Dataset

Data have been collected from Baghdad's hospitals, diagnostic centers, and clinic centers. Most of the data have been collected from hospital-admitted patients. Patients are questioned, test results are examined, and essential

attribute data are collected. The dataset has 59 patients' test results and responses to various questions. Table 1 summarizes the specifications of the 19 attributes (18 independent, while the target attribute is dependent) that comprise the dataset. Differences in the ranges of variables appear; hence, it is essential to normalize the data. Normalization is accomplished according to the equation as follows

$$x_{normalised} = \frac{x - \min x}{\max(x) - \min(x)} \quad (5)$$

where $X_{normalized}$ is the updated normalized value. The lowest value of each feature is taken as '0', the highest value is considered as '1', and all other values are converted to an integer between '0' and '1'.

Table 1: Dataset Description

Attribute	Type	Value
Age	Nominal	26–90
Gender	Nominal	Male, Female
Occupation	Nominal	Businessman, Driver, Farmer, Worker, Employed, Unemployed, Housewife, Service-holder.
Family history	Nominal	Yes, No
smoking	Nominal	Yes, No
Obesity	Nominal	Yes, No
Diet	Nominal	Normal, Abnormal
Physical activity	Nominal	Yes, No
Stress	Nominal	Normal, Abnormal
Chest pain type	Nominal	Typical angina, Atypical angina, Non-angina pain
Previous chest pain	Nominal	Yes, No
Edema	Nominal	Yes, No
Blood pressure systolic (bp_systolic)	Nominal	80–180
Blood pressure diastolic (bp_diastolic)	Nominal	60–120
Heart rate	Nominal	60–120
Diabetes	Nominal	Yes, No
Troponin	Nominal	Positive, Negative
ECG	Nominal	Normal, Abnormal
Target	Nominal	Class 1 (Patient has heart disease) and Class 0 (Patient has no heart disease)

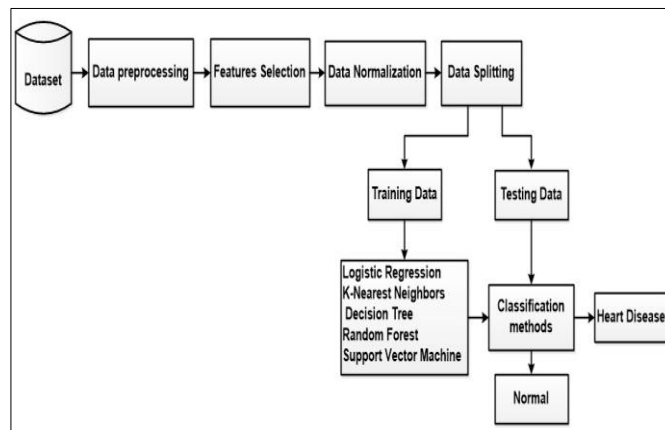


Figure 7. Frame of the proposed method

5.2 Data pre-processing

Initially, data undergo pre-processing immediately after collection, involving several tasks. The collected data consists of numerous records that may contain missing data or values, especially in features like age. Typically, missing values are substituted with the nearest or most proximate values within their respective features. The

values of the target label in a classification model are then transformed into a non-numeric form. Subsequently, the dataset is divided into two groups: training and testing. The data is prepared for model classification. Once the data pre-processing is concluded, the model selection stage commences for classification. Normalization becomes imperative due to the substantial variation in the range of dataset values, particularly following the encoding of nominal values into real numbers 1 and 2. This scaling is crucial to prevent more extensive numeric ranges from overpowering attributes with smaller numeric ranges. Additionally, normalization accelerates the execution of algorithms by circumventing the utilization of an extensive range of numbers[22]. The data is transformed to fit within the (0–1) interval, as per equation (1), where x denotes the original value of the attribute, x -normalized represents the scaled value, $\min a$ signifies the minimum value of attribute a , and $\max a$ means the maximum value of the attribute a

$$x_{normalised} = \frac{x - \min a}{\max a - \min a} \quad (6)$$

5.3 Feature Selection

Identifying crucial features related to cardiovascular disease disorders is an integral component in predicting chronic cardiovascular disease. In this phase, various features, including age and gender, which signify each patient's personal information, are chosen. Additionally, various clinical features obtained from diverse medical tests are included in the selection process.[23].

5.4 Data classification

Classification plays a crucial role in data mining, assigning the collected items to specific target classes or categories. The primary goal of classification is to predict the target class for all case data accurately. Following data pre-processing, features are integrated into a classification model. Several well-known classification models, including Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression (LR), and K-Nearest Neighbors (K-NN), are commonly utilized in this process.

5.5 Prediction

In this phase, the chosen features are mapped onto the training model to classify the given features and predict the presence or absence of cardiovascular disease. A specialist doctor labels the collected cardiovascular disease dataset to generate predictions. The classification is structured as a multi-class problem, categorizing medical data into different classes, each corresponding to a specific category of cardiovascular disease. Based on the selected essential features, this process can determine the probability of a patient suffering from cardiovascular disease [24].

6. Result and discussion

The MATLAB R2019a tool was used to conduct simulations, utilizing a dataset from the UCI repository[25, 26]. The dataset comprises 1385 samples, with 969 (constituting 70% of the dataset) allocated for training the model. The remaining 416 samples (30% of the dataset) are reserved for validation. Notably, the training set of 969 samples encompasses representatives from all classes, ensuring a balanced inclusion of 70% of each class in the training phase. The confusion matrix of Logistic Regression, K-nearest neighbor, Decision Tree, Random Forest, and Support Vector Machine is represented in Table 2 and Figure 8. Figure 8 displays the outcomes and accuracies the implemented machine learning algorithms achieved, presenting the average accuracy across diverse trained datasets. The models were trained with these subsets, and the overall accuracy was computed. Random Forest (RF) exhibited superior performance, achieving the highest average accuracy at 87.76%. Figure 9 illustrates the variation in accuracy with different training dataset sizes, encompassing 50% to 90%. Conversely, the other algorithms exhibit excellent performance, reaching peak accuracy values when trained with a dataset size of 80%.

Table 2: Confusion Matrix Parameters

Method	TP	FP	FN	TN
--------	----	----	----	----

LR	35	2	6	19
K-NN	34	1	5	20
SVM	34	2	5	20
DT	32	4	4	21
RF	34	2	2	23

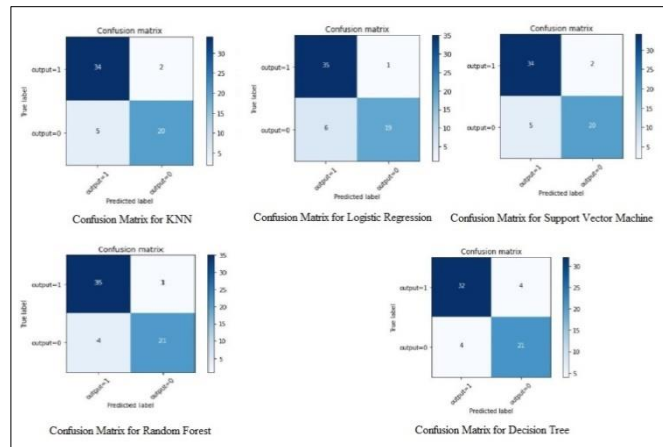


Figure 8. Confusion matrix of different machine learning models

80% identified the training size as the optimal range for predicting the ailment and obtaining accuracy, precision, recall, and F1 score at this size to showcase the maximum peak value. A 50% split poses a higher risk of underfitting, while a 90% split may lead to overfitting. Utilizing an 80% split helps overcome this scenario. Many researchers have employed an 80% split to train their models[2, 27]. Table 3 and Figs. 9,10,11,12 present the accuracy of the algorithm used in the research. Classification models such as Logistic Regression (LR), K-nearest neighbor (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) are used to analyze and predict heart attack risks, and the accuracy of each model is compared. The best classification model is recognized by the accuracy percentage found from the experiment. The LR algorithm's accuracy, precision, recall, and F1 scores were 79.49%, 82.90%, 93.50%, and 87.88%, respectively. K-NN demonstrated an accuracy, precision, recall, and F1 score of 81.52%, 82.60%, 96.80%, and 89.11%, respectively. SVM achieved an accuracy, precision, recall, and F1 score of 81.20%, 81.40%, 98.90%, and 89.32%, respectively. The DT exhibited an accuracy, precision, recall, and F1 score of 80.41%, 79.50%, 100.00%, and 88.57%, respectively. RF showcased an accuracy, precision, recall, and F1 score of 87.76%, 87.00%, 93.50%, and 90.16%, respectively. RF yielded satisfactory results among the applied algorithms, demonstrating extraordinary performance across various statistical metrics. Table 3 and Figure 9 illustrate the performance of the proposed algorithms. RF displayed the highest score at 87.76%, while LR exhibited the lowest score at 79.49% among the algorithms.

Table 3 shows the results with an 80% train size

Classification methods	Accuracy	Precision	Recall	F1_score
LR	79.49%	82.90%	93.50%	87.88%
K-NN	81.52%	82.60%	96.80%	89.11%
SVM	81.20%	81.40%	98.90%	89.32%
DT	80.41%	79.50%	100.00%	88.57%
RF	87.76%	87.00%	93.50%	90.16%

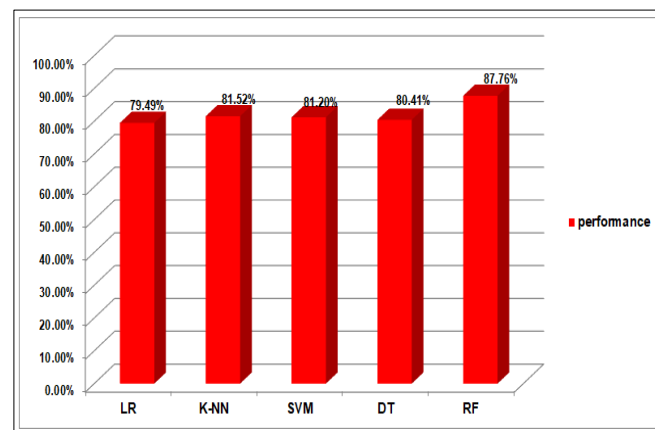


Figure.9. Comparison of accuracy for various algorithms for the prediction of heart diseases

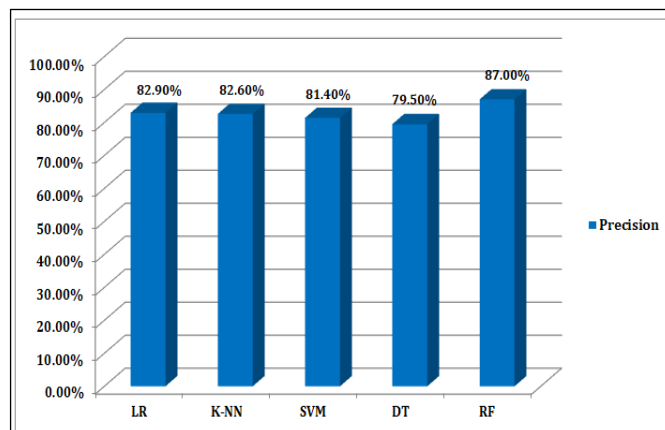


Figure 10. Comparison of precision for various algorithms for the prediction of heart diseases

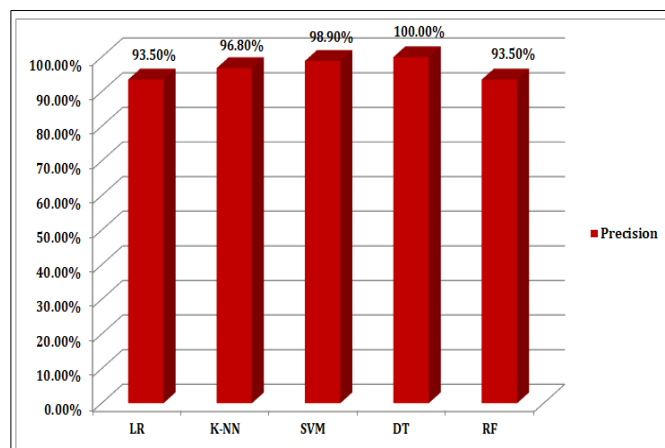


Figure 11. Comparison of precision of various algorithms for the Recall of heart diseases

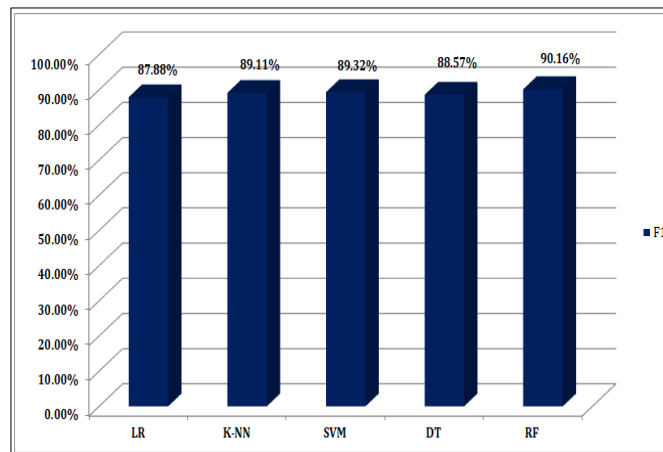


Figure 12. Comparison of precision of various algorithms for the F1 of heart diseases

Compared with other studies, various algorithms were employed, including LR, SVM, K-NN, RF, DT, and others, achieving a notable accuracy score of 81%. This research demonstrated satisfactory performance using the same dataset as a previous study. Among the applied algorithms, RF exhibited significant efficacy, reaching an accuracy of 87.76%. Nevertheless, the remaining algorithms displayed moderate performance in the early-stage prognosis of cardiovascular disease. Compared to other researchers utilizing machine learning techniques such as LR, SVM, K-NN, RF, DT, and more, our experiment results showcase a commendable performance across various algorithms. Notably, applying these algorithms in our study yielded promising outcomes, with an overall accuracy that stands out at a competitive level. In particular, Random Forest (RF) demonstrated significant efficacy, achieving an accuracy rate of 87.76%, surpassing the performance observed in other studies, 71.87[28]. This indicates the robustness and effectiveness of our approach in the context of predicting cardiovascular disease. In addition, LR and SVM established noteworthy efficacy, achieving an accuracy rate of 79.49% and 81.20%, respectively, surpassing the performance observed in other studies, which is 73.39%, 71%, 71.24% [29, 30]. Moreover, in our research, Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) exhibited noteworthy efficacy, achieving accuracy rates of 87.76%, 79.49%, and 81.20%, respectively. These results surpassed the performance observed in other studies, where accuracy rates were reported at 77%, 81%, 71.24%, and 77% [31-33]. The heightened accuracy achieved by LR and SVM in our experiment suggests the effectiveness of these algorithms in predicting cardiovascular disease compared to the outcomes reported in previous research. This underscores the robustness and reliability of our approach, emphasizing the potential for improved accuracy and performance in applying machine learning techniques for cardiovascular disease prognosis. The results suggest that our experiment contributes valuable insights and competitive outcomes in using machine learning algorithms for cardiovascular disease prognosis, showcasing the potential of our methodology in this research domain.

7. Conclusion

This paper presents various prediction algorithms for the early-stage prediction and diagnosis of cardiovascular disease. This study evaluates the performance of selected Machine Learning (ML) algorithms through classification measures to determine the presence of hepatitis based on various independent attributes. There is a massive amount of data when considering the domain related to health and medicine, which remains the same when referring to the records associated with detecting heart-related disease. Therefore, this type of record increases the complexity of determining the conclusions made by medical practitioners. To have this process done with ease, we use specific ML-based models to compute the data and give accurate outcomes by providing proper predictions for diseases related to the heart. The different techniques used are mentioned and briefly explained in this paper. The accuracy of each of the algorithms is measured and compared to determine which ML techniques

will possess the most potential in giving correct predictions. When considering false negatives, the Logistic regression and the RF technique play a vital role. Still, when considering the accuracy score as the evaluative method, the RF algorithm remains the best, with 87.762%. This helps appropriately implement the model and guides the use of the machine-learning technique. The future work of this research will likely focus on improving the accuracy and effectiveness of machine learning techniques for heart disease analysis and prediction, with the ultimate goal of improving patient outcomes and reducing the burden of cardiovascular disease on society.

Authors' Contributions: All authors have worked equally to write and review the manuscript.

Data Availability Statement: All data are presented in the paper, along with their related references, in the Application section.

Conflicts of Interest: The authors declare no conflict of interest.

Funding: There are no funds received for this paper.

Use of Generative-AI tools declaration: The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

References

- [1] W. Sun, P. Zhang, Z. Wang, and D. Li, "Prediction of cardiovascular diseases based on machine learning," *ASP Transactions on Internet of Things*, vol. 1, pp. 30-35, 2021.
- [2] A. S. Alenizi and K. A. Al-karawi, "Machine learning approach for diabetes prediction," in *International Congress on Information and Communication Technology*, 2023, pp. 745-756.
- [3] M. M. Abdelwahab, K. A. Al-Karawi, and H. Smary, "Integrating gene selection and deep learning for enhanced Autisms' disease prediction: a comparative study using microarray data," *AIMS Mathematics*, vol. 9, pp. 17827–17846, 2024.
- [4] S. S. Bhuyan and A. K. Mishra, "Machine Learning Algorithms for Heart Disease Prediction," *Asian Journal For Convergence In Technology (AJCT) ISSN-2350-1146*, vol. 8, pp. 87-91, 2022.
- [5] M. M. Abdelwahab, K. A. Al-Karawi, and H. E. Smary, "Deep learning-based prediction of Alzheimer's disease using microarray gene expression data," *Biomedicines*, vol. 11, p. 3304, 2023.
- [6] WHO. (2012). *World health Organization, The top 10 causes of death.* available at: . Available: <http://www.who.int/mediacentre/factsheets/fs310/en/>
- [7] J. Pirgazi, A. Ghanbari Sorkhi, and M. Iranpour Mobarkeh, "An Accurate Heart Disease Prognosis Using Machine Intelligence and IoMT," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.
- [8] M. Sarveshvar, A. Gogoi, A. K. Chaubey, S. Rohit, and T. Mahesh, "Performance of different machine learning techniques for the prediction of heart diseases," in *2021 international conference on forensics, analytics, big data, security (FABS)*, 2021, pp. 1-4.
- [9] H. Smary, K. A. Al-Karawi, and M. M. Abdelwahab, "Using Voice Technologies to Support Disabled People," *Journal of Disability Research*, vol. 3, p. 20230063, 2024.
- [10] K. A. Al-Karawi, "Face mask effects on speaker verification performance in the presence of noise," *Multimedia Tools and Applications*, pp. 1-14, 2023.
- [11] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network," *Ieee Access*, vol. 7, pp. 34938-34945, 2019.
- [12] A. S. Alenizi and K. A. Al-Karawi, "Internet of Things (IoT) Adoption: Challenges and Barriers," in *Proceedings of Seventh International Congress on Information and Communication Technology*, 2023, pp. 217-229.
- [13] X. Yuan, J. Chen, K. Zhang, Y. Wu, and T. Yang, "A stable AI-based binary and multiple class heart disease prediction model for IoMT," *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 2032-2040, 2021.
- [14] E. Sogancioglu, K. Murphy, E. Calli, E. T. Scholten, S. Schalekamp, and B. Van Ginneken, "Cardiomegaly detection on chest radiographs: Segmentation versus classification," *IEEE Access*, vol. 8, pp. 94631-94642, 2020.
- [15] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," in *The Impact of Digital Technologies on Public Health in Developed and Developing Countries: 18th International Conference, ICOST 2020, Hammamet, Tunisia, June 24–26, 2020, Proceedings 18*, 2020, pp. 299-306.
- [16] A. Tyagi and R. Mehra, "Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with GOA," *SN Applied Sciences*, vol. 3, pp. 1-14, 2021.

- [17] A. Palmer, R. Jiménez, and E. Gervilla, "Data mining: Machine learning and statistical techniques," *Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.)*, pp. 373-396, 2011.
- [18] K. A. Al-Karawi and D. Y. Mohammed, "Using combined features to improve speaker verification in the face of limited reverberant data," *International Journal of Speech Technology*, pp. 1-11, 2023.
- [19] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, pp. 2038-2048, 2007.
- [20] N. Nadar and R. Kamatchi, "Information and communication-based collaborative learning and behavior modeling using machine learning algorithm," in *Social Media and Machine Learning*, ed: IntechOpen, 2019.
- [21] S. Grampurohit and C. Sagarnal, "Disease prediction using machine learning algorithms," in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-7.
- [22] J.-M. Jo, "Effectiveness of normalization pre-processing of big data to the machine learning performance," *The Journal of the Korea institute of electronic communication sciences*, vol. 14, pp. 547-552, 2019.
- [23] D. Paper, *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*: Apress, 2019.
- [24] S. Kohara, T. Tateyama, A. H. Foruzan, A. Furukawa, S. Kanasaki, M. Wakamiya, *et al.*, "Application of statistical shape model to diagnosis of liver disease," in *The 2nd International Conference on Software Engineering and Data Mining*, 2010, pp. 680-683.
- [25] M. B. Butt, M. Alfayad, S. Saqib, M. Khan, M. Ahmad, M. A. Khan, *et al.*, "Diagnosing the stage of hepatitis C using machine learning," *Journal of Healthcare Engineering*, vol. 2021, 2021.
- [26] M. B. Butt, M. Alfayad, S. Saqib, M. Khan, M. Ahmad, M. A. Khan, *et al.*, "Research Article Diagnosing the Stage of Hepatitis C Using Machine Learning," 2021.
- [27] M. M. S. Raihan, A. B. Shams, and R. B. Preo, "Multi-class electrogastrogram (EGG) signal classification using machine learning algorithms," in *2020 23rd International conference on computer and information technology (ICCIT)*, 2020, pp. 1-6.
- [28] J. Singh, S. Bagga, and R. Kaur, "Software-based prediction of liver disease with feature selection and classification techniques," *Procedia Computer Science*, vol. 167, pp. 1970-1980, 2020.
- [29] S. Sontakke, J. Lohokare, and R. Dani, "Diagnosis of liver diseases using machine learning," in *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 2017, pp. 129-133.
- [30] S. Bahramirad, A. Mustapha, and M. Eshraghi, "Classification of liver disease diagnosis: A comparative study," in *2013 Second International Conference on Informatics & Applications (ICIA)*, 2013, pp. 42-46.
- [31] L. A. Auxilia, "Accuracy prediction using machine learning techniques for indian patient liver disease," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 45-50.
- [32] A. S. Singh, M. Irfan, and A. Chowdhury, "Prediction of liver disease using classification algorithms," in *2018 4th international conference on computing communication and automation (ICCCA)*, 2018, pp. 1-3.
- [33] N. Nahar, F. Ara, M. A. I. Neloy, V. Barua, M. S. Hossain, and K. Andersson, "A comparative analysis of the ensemble method for liver disease prediction," in *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, 2019, pp. 1-6.